Biostatistics (2000), **1**, 3, *pp.* 231–246
Printed in Great Britain

On meta-analytic assessment of surrogate outcomes

MITCHELL H. GAIL*
National Cancer Institute, Division of Cancer Epidemiology and Genetics,
Executive Plaza South, Room 8032, 1620 Executive Boulevard, MSC 7244, Bethesda,
MD 20892-7244, USA
gailm@exchange.nih.gov

RUTH PFEIFFER
National Cancer Institute, Division of Cancer Epidemiology and Genetics,
Executive Plaza South, Room 8017, 1620 Executive Boulevard,
MSC 7244, Bethesda, MD 20892-7244, USA

HANS C. VAN HOUWELINGEN
Medical Statistics, Leiden University Medical Center, Wassenaarseweg 62,
PO Box 9604, 2300 RC Leiden, The Netherlands

RAYMOND J. CARROLL
Department of Statistics, Texas A&M University, College Station,
TAMU 3143, TX 77843-3143, USA

SUMMARY

We discuss the strengths and weaknesses of the meta-analytic approach to estimating the effect of a new treatment on a true clinical outcome measure, T, from the effect of treatment on a surrogate response, S. The meta-analytic approach (see Daniels and Hughes, 1997) uses data from a series of previous studies of interventions similar to the new treatment. The data are used to estimate relationships between summary measures of treatment effects on T and S that can be used to infer the magnitude of the effect of the new treatment on T from its effects on S. We extend the class of models to cover a broad range of applications in which the parameters define features of the marginal distribution of (T, S). We present a new bootstrap procedure to allow for the variability in estimating the distribution that governs the between-study variation. Ignoring this variability can lead to confidence intervals that are much too narrow. The meta-analytic approach relies on quite different data and assumptions than procedures that depend, for example, on the conditional independence, at the individual level, of treatment and T, given S (see Prentice, 1989). Meta-analytic calculations in this paper can be used to determine whether a new study, based only on S, will yield estimates of the treatment effect on T that are precise enough to be useful. Compared to direct measurement on T, the meta-analytic approach has a number of limitations, including likely serious loss of precision and difficulties in defining the class of previous studies to be used to predict the effects on T for a new intervention.

Keywords: Bootstrap; Bootstrap confidence intervals; Clinical trials; Empirical Bayes procedures; Meta-analysis; Surrogate endpoints.

*To whom correspondence should be addressed

Published by Oxford University Press

1. INTRODUCTION

There is great interest in using surrogate endpoints S, in clinical trials, in place of clinically relevant main ('true') endpoints T. Surrogate endpoints may yield response data earlier than main endpoints, and may require smaller sample sizes. Reliance on surrogates is problematic, however, because it is difficult to gauge how reliably one can infer effects of treatment on T from data on S.

Much work on surrogate markers attempts to relate an individual's true response to treatment to that individual's surrogate response to treatment. Such research has defined conditions under which measurements on S may be reliably used to test for a treatment effect on the main endpoint, T (Prentice, 1989; Buyse and Molenberghs, 1998). An essential condition is that T be conditionally independent of treatment given S. Because strict conditional independence may hold only infrequently and is difficult to confirm, others have considered the related estimation of the 'percentage of treatment effect explained' by the surrogate (Freedman *et al.*, 1992; Lin *et al.*, 1997).

This literature on the usefulness of a surrogate for evaluating treatments for the individuals who participate in a given study has two limitations. First, there has been comparatively little discussion of how data on S can be used to estimate the magnitude of the effects of treatment on T. Second, even if in a particular study of a given drug, data on both T and S confirm that S can be used to predict T reliably for each individual in the study, whether on the new drug or on the control treatment, it does not follow that S will be reliable for testing for a treatment effect or for estimating the magnitude of the treatment effect for another drug in another study. Empirical evidence on these points can derive from a series of studies on drugs of a given type. This is the rationale for a meta-analytic evaluation of surrogate markers.

Two recent papers used a meta-analytic approach to estimate effects of treatment, Z, on T from data on S. The idea is that one can 'borrow information' from previous similar studies on the relationships between T and S in treated ($Z = 1$) and control ($Z = 2$) groups. Daniels and Hughes (1997) regressed treatment effects for T on treatment effects for S in a meta-analysis of previous studies on the effects of anti-retroviral agents, and they used information on the regression relationships to estimate the treatment effect on T in a new study from an estimate of the treatment effect on S in the new study. In recent work, Buyse *et al.* (2000) (BMBRG) proposed a linear mixed model for conducting such a meta-analysis. This model allows one to estimate treatment effects on T as linear functions of the separate values of S in treated and untreated groups, rather than simply as a linear function of the estimated treatment effect on S. We discuss a multivariate normal model closely related to that of BMBRG but with a more general covariance structure (Section 3.1). The generality can be important if treatment affects not only population means but also variances or covariances, and for more complex problems (see Sections 2, 3.2, and 3.3).

As pointed out by Daniels and Hughes (1997), it can be difficult to specify a realistic joint distribution for T and S given Z. For example, analysts may not agree on the joint distribution of T and S for time-to-response data. For this reason, we introduce separate marginal models for T and for S in treated and in untreated groups (Section 2). We do not even require that the parameters we estimate completely define these marginal distributions but only that they characterize important features of the responses to T and S that can be used to estimate treatment effects on T. These models can accommodate complex measurements, such as piecewise exponential survival data or repeated measurements (Section 3.3).

In Section 2, we present notation, a general meta-analytic sampling framework and a general formulation. Applications to the normal model (Section 3.1), dichotomous outcomes (Section 3.2), and a general marginal model for survival (Section 3.3) follow. In Section 4 we discuss the impact of uncertainty in parameter estimates on the precision of prediction intervals for the treatment effects on the true endpoint. Data from the REGRESS trial (Jukema *et al.*, 1995) are analyzed in Section 5 to illustrate the method. We defer a discussion of some of the serious practical and theoretical limitations of the meta-analytic approach to Section 6.

2. Notation, sampling framework, and general formulation

We seek to estimate treatment effects on T for a new drug using only incomplete data on S in the new study (N) and using information from K previous 'similar' studies with complete data on T and S. A crucial assumption is that we can identify a class C of similar drug studies to which the current drug study (N) is related. The ith drug study in class C is regarded as a random sample from members of this class. The ith study has parameters $\theta_i = (\theta_{1Ti}^{\mathrm{T}}, \theta_{1Si}^{\mathrm{T}}, \theta_{2Ti}^{\mathrm{T}}, \theta_{2Si}^{\mathrm{T}})^{\mathrm{T}}$, where $(\theta_{1Ti}, \theta_{2Ti})$ are features of the marginal distribution of T in the experimental ($Z = 1$) and control groups ($Z = 2$), respectively, and $(\theta_{1Si}, \theta_{2Si})$ are likewise features of the marginal distribution of S for $Z = 1$ and $Z = 2$, respectively. The treatment effect, $\delta_i = \delta(\theta_{1Ti}, \theta_{2Ti})$ is a function of $(\theta_{1Ti}, \theta_{2Ti})$. We hope that knowledge of $\theta_{SN} = (\theta_{1SN}^{\mathrm{T}}, \theta_{2SN}^{\mathrm{T}})^{\mathrm{T}}$ obtained in experiment N will yield information on $\theta_{TN} = (\theta_{1TN}^{\mathrm{T}}, \theta_{2TN}^{\mathrm{T}})^{\mathrm{T}}$, and hence on δ_N.

We assume the θ_i are drawn at random from a multivariate normal distribution with mean μ and covariance matrix ϕ (Figure 1). Even if the same drug is tested in two different populations, the values of θ_i may differ, because populations may differ in levels of risk in the absence of treatment and because they may differ in responsiveness to treatment due to variations in compliance or treatment-covariate interactions. Therefore, in the formulations that follow, we regard the 'drug study' as the element of the class, rather than the drug alone.

The objective is to use the conditional distribution of θ_N given the data on S in the treated and untreated groups in study N, and given knowledge of μ and ϕ obtained from previous studies, to estimate δ_N and obtain confidence intervals for it.

The basic data and analytical approach are outlined in Figures 1 and 2. Let (T_{1ij}, S_{1ij}) be the response of subject j in trial i in treatment $Z = 1$, for $i = 1, \ldots, K$ and $j = 1, \ldots, n_i$, and define (T_{2ij}, S_{2ij}) similarly for $Z = 2$, $i = 1, \ldots, K$ and $j = 1, \ldots, m_i$. The surrogate responses may be vectors, but we use scalar notation for simplicity. For $z = 1, 2$, the distribution of (T_{zij}, S_{zij}) depends on $\theta_{zi} = (\theta_{zTi}^{\mathrm{T}}, \theta_{zSi}^{\mathrm{T}})^{\mathrm{T}}$ and possibly on other parameters γ_i. As indicated in Figure 1, data from experiment i are used to estimate $\theta_i = (\theta_{1i}^{\mathrm{T}}, \theta_{2i}^{\mathrm{T}})^T$ and the conditional covariance matrix $\mathrm{cov}(\widehat{\theta_i}|\theta_i) = \Sigma_i$. Because $(\widehat{\theta}_{1i}, \widehat{\theta}_{2i})$ are conditionally independent given θ_i, $\Sigma_i = \mathrm{diag}(\Sigma_{11i}, \Sigma_{22i})$ is block-diagonal. We call σ_{22i} the submatrix of Σ_{11i} corresponding to $\widehat{\theta}_{1Si}$ and σ_{44i} the submatrix of Σ_{22i} corresponding to $\widehat{\theta}_{2Si}$.

We assume that the estimate $\widehat{\theta_i}$ is obtained from estimating equations $U_{1Ti}(\theta_{1Ti}) = \sum_{j=1}^{n_i} U_{1Tij}(\theta_{1Ti}) = 0$, $U_{1Si}(\theta_{1Si}) = \sum_{j=1}^{n_i} U_{1Sij}(\theta_{1Si}) = 0$, $U_{2Ti}(\theta_{2Ti}) = \sum_{j=1}^{m_i} U_{2Tij}(\theta_{2Ti}) = 0$, $U_{2Si}(\theta_{2Si}) = \sum_{j=1}^{m_i} U_{2Sij}(\theta_{2Si}) = 0$. Note that $U_{1Ti}(\cdot)$ is not a function of $(\theta_{1Si}, \theta_{2Ti}, \theta_{2Si}, \gamma_i)$, and other estimating equations are likewise functionally independent of parameters not in their argument. We assume sufficient regularity conditions such that, conditional on θ_i and γ_i, $\widehat{\theta_i}$ is normally distributed with covariance matrix Σ_i. Unconditionally, $\widehat{\theta_i}$ is normally distributed with mean μ and covariance matrix $\phi + \Sigma_i$. The covariance matrix Σ_i can be estimated as a 'sandwich', $\hat{\Sigma}_{11i} = n_i^{-1}(n_i - p - q)^{-1}\hat{B}_1 V \hat{B}_1^{\mathrm{T}}$, with the entries of V given by $V_{11} = \sum_{j=1}^{n_i} U_{1Tij}(\hat{\theta}_{1Ti})U_{1Tij}^{\mathrm{T}}(\hat{\theta}_{1Ti})$, $V_{12} = V_{21} = \sum_{j=1}^{n_i} U_{1Tij}(\hat{\theta}_{1Ti})U_{1Sij}^{\mathrm{T}}(\hat{\theta}_{1Si})$ and $V_{22} = \sum_{j=1}^{n_i} U_{1Sij}(\hat{\theta}_{1Si})U_{1Sij}^{\mathrm{T}}(\hat{\theta}_{1Si})$ (see e.g. Appendix B in Carroll *et al.*, 1995). Here p and q are the dimensions of θ_{1Ti} and θ_{1Si} respectively, and $B_1^{-1} = n_i \mathrm{diag}\left\{E\left(\frac{\partial U_{1Tij}}{\partial \theta_{1Ti}}\right)^{\mathrm{T}}, E\left(\frac{\partial U_{1Sij}}{\partial \theta_{1Si}}\right)^{\mathrm{T}}\right\}^{\mathrm{T}}$, which can be estimated by $\left\{\sum_{j=1}^{n_i}\left(\frac{\partial U_{1Tij}}{\partial \theta_{1Ti}}\right)^{\mathrm{T}}, \sum_{j=1}^{n_i}\left(\frac{\partial U_{1Sij}}{\partial \theta_{1Si}}\right)^{\mathrm{T}}\right\}^{\mathrm{T}}$. A similar sandwich estimate is available for Σ_{22i}. For example, in the simple multivariate normal case (Section 3.1), $U_{1Tij}(\theta_{1Ti}) = T_{1ij} - \theta_{1ti}$, $U_{1Sij}(\theta_{1Si}) = S_{1ij} - \theta_{1Si}$, $B_1^{-1} = -n_i$ diag $(I_{p \times p}, I_{q \times q})$, and the estimate of $\mathrm{cov}(\hat{\theta}_{1Ti}, \hat{\theta}_{1Si})$ is $n_i^{-1}\{n_i - p - q\}^{-1} \sum_{j=1}^{n_i}(T_{1ij} - T_{1i})(S_{1ij} - S_{1i})^{\mathrm{T}}$, where T_{1i} and S_{1i} are averages over j of T_{1ij} and S_{1ij}.

Now consider the new experiment, N. From data on S (Figure 1), we obtain estimates $(\widehat{\theta}_{1SN}, \widehat{\theta}_{2SN}, \widehat{\sigma}_{22N}, \widehat{\sigma}_{44N})$. Let D and W be defined so that $\theta_{TN} = D\theta_N$ and $\theta_{SN} = W\theta_N$. The $2p \times 2(p + q)$ matrix D has ones for the elements $(1, 1), \ldots, (p, p)$ and the elements $(p + 1, p + q + 1)$, $(p + 2, p +$

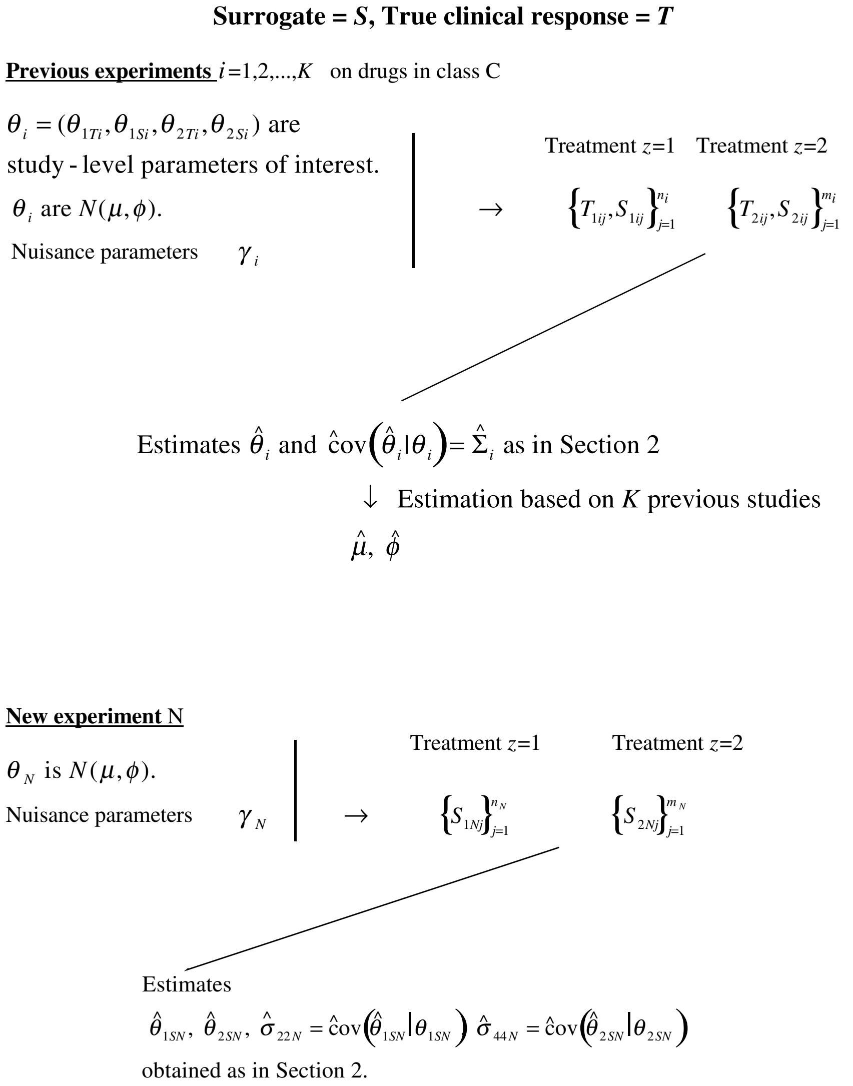
Surrogate = S, True clinical response = T
Previous experiments $i = 1,2,...,K$ on drugs in class C
$\theta_i = (\theta_{1Ti}, \theta_{1Si}, \theta_{2Ti}, \theta_{2Si})$ are study - level parameters of interest.
θ_i are $N(\mu, \phi)$.
Nuisance parameters γ_i
Treatment z=1
Treatment z=2
$\rightarrow$
$\{T_{1ij}, S_{1ij}\}_{j=1}^{n_i}$
$\{T_{2ij}, S_{2ij}\}_{j=1}^{m_i}$
Estimates $\hat{\theta}_i$ and $\hat{\text{cov}}(\hat{\theta}_i | \theta_i) = \hat{\Sigma}_i$ as in Section 2
$\downarrow$ Estimation based on K previous studies
$\hat{\mu}$, $\hat{\phi}$
New experiment N
θ_N is $N(\mu, \phi)$.
Nuisance parameters γ_N
Treatment z=1
Treatment z=2
$\rightarrow$
$\{S_{1Nj}\}_{j=1}^{n_N}$
$\{S_{2Nj}\}_{j=1}^{m_N}$
Estimates
$\hat{\theta}_{1SN}$, $\hat{\theta}_{2SN}$, $\hat{\sigma}_{22N} = \hat{\text{cov}}(\hat{\theta}_{1SN} | \theta_{1SN})$ $\hat{\sigma}_{44N} = \hat{\text{cov}}(\hat{\theta}_{2SN} | \theta_{2SN})$
obtained as in Section 2.

Fig. 1. Description of parameters, data, and estimates for the K experiments with data on true and surrogate outcomes and for the new experiment, N, with data only on surrogates.

$q + 2), \ldots, (2p, 2p + q)$ and zeros elsewhere. The $2q \times 2(p + q)$ matrix W has ones for the elements $(1, p + 1), (2, p + 2), \ldots, (q, p + q)$ and the elements $(q + 1, 2p + q + 1), \ldots, (2q, 2p + 2q)$ and zeros elsewhere. Because $(\theta_N, \widehat{\theta}_N)$ are jointly normally distributed, with mean $(\mu^{\mathrm{T}}, \mu^{\mathrm{T}})^{\mathrm{T}}$, variances ϕ and $\phi + \Sigma_N$ and covariance ϕ, $(\theta_{TN}^{\mathrm{T}}, \widehat{\theta}_{SN}^{\mathrm{T}})^{\mathrm{T}}$ is normally distributed with mean $\{(D\mu)^{\mathrm{T}}, (W\mu)^{\mathrm{T}}\}^{\mathrm{T}}$, and with covariance matrix defined by $\text{cov}(\theta_{TN}) = D\phi D^{\mathrm{T}}$, $\text{cov}(\widehat{\theta}_{SN}) = W(\phi + \Sigma_N)W^{\mathrm{T}}$, and $\text{cov}(\theta_{TN}, \widehat{\theta}_{SN}) = D\phi W^{\mathrm{T}}$. Hence, the conditional distribution of θ_{TN} given $\widehat{\theta}_{SN}$ is normal with mean and covariance matrix given by:

$$E(\theta_{TN} | \widehat{\theta}_{SN}) = D\mu + D\phi W^{\mathrm{T}}\{W(\phi + \Sigma_N)W^{\mathrm{T}}\}^{-1}(\widehat{\theta}_{SN} - W\mu); \tag{1}$$

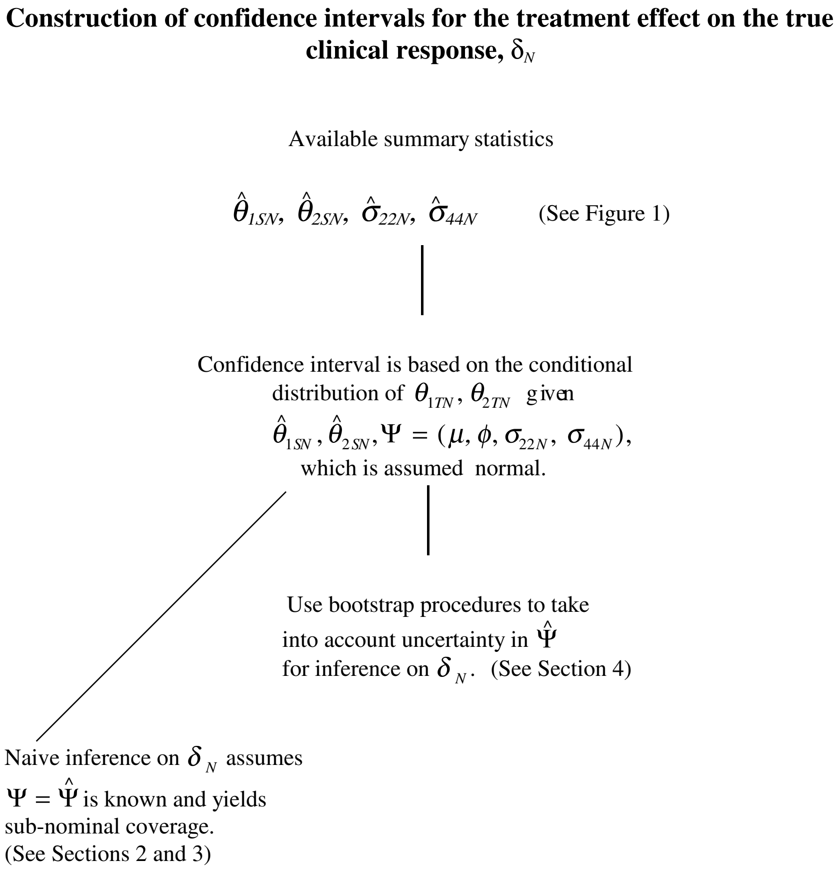
Construction of confidence intervals for the treatment effect on the true clinical response, δ_N
Available summary statistics
$\hat{\theta}_{1SN}, \hat{\theta}_{2SN}, \hat{\sigma}_{22N}, \hat{\sigma}_{44N}$ (See Figure 1)
Confidence interval is based on the conditional distribution of $\theta_{1TN}, \theta_{2TN}$ given $\hat{\theta}_{1SN}, \hat{\theta}_{2SN}, \Psi = (\mu, \phi, \sigma_{22N}, \sigma_{44N})$, which is assumed normal.
Use bootstrap procedures to take into account uncertainty in $\hat{\Psi}$ for inference on δ_N. (See Section 4)
Naive inference on δ_N assumes $\Psi = \hat{\Psi}$ is known and yields sub-nominal coverage. (See Sections 2 and 3)

Fig. 2. Inference for the true treatment effect, δ_N, in the new experiment based on surrogate data in the new experiment and knowledge of Ψ.

$$\mathrm{cov}(\theta_{TN}|\widehat{\theta}_{SN}) = D\phi D^{\mathrm{T}} - D\phi W^{\mathrm{T}}\{W(\phi + \Sigma_N)W^{\mathrm{T}}\}^{-1}W\phi D^{\mathrm{T}}. \tag{2}$$

Note that (2) only depends on Σ_N through the elements $(\sigma_{22N}, \sigma_{44N})$ corresponding to $(\widehat{\theta}_{1SN}, \widehat{\theta}_{2SN})$.

If $\Psi = (\mu, \phi, \sigma_{22N}, \sigma_{44N})$ were known, then we could base inference about the treatment effect $\delta_N = \delta(\theta_{TN})$ on equations (1) and (2) (see Figure 2). For example, if θ_{1TN} and θ_{2TN} are scalars and $\delta_N = \theta_{1TN} - \theta_{2TN} = R\theta_{TN}$, where $R = (1, -1)$, then the conditional distribution of δ_N is normal with mean and variance given by:

$$M(\Psi) = RE(\theta_{TN}|\widehat{\theta}_{SN}); \tag{3}$$

$$V(\Psi) = R\mathrm{cov}(\theta_{TN}|\widehat{\theta}_{SN})R^{\mathrm{T}}. \tag{4}$$

A 95% confidence interval for δ_N would then be $M(\Psi) \pm 1.96V^{1/2}(\Psi)$. If δ_N were a nonlinear function of the elements of θ_{TN}, its conditional distribution could still be determined analytically in some cases, or more generally by simulating values of θ_{TN} from the conditional distribution based on (1) and (2).

In practice, Ψ is unknown and must be estimated. We can estimate the block-diagonal covariance matrix $W\Sigma_N W^{\mathrm{T}} = \mathrm{diag}(\sigma_{22N}, \sigma_{44N})$ from individual-level study data, (Figure 1). To estimate (μ, ϕ), we rely on data from the studies $i = 1, \ldots, K$. Because $\widehat{\theta_i}$ has mean μ and covariance matrix $\phi + \Sigma_i$, and because Σ_i can be estimated from individual-level study data, various methods such as maximum likelihood, restricted maximum likelihood or empirical Bayes can be used to estimate (μ, ϕ) (Figure 1).

Estimation of Ψ means that the confidence intervals for δ_N described above will usually have subnominal coverage (Figure 2). In Section 4, we describe bootstrap methods to adjust the intervals for estimation of Ψ.

3. Special cases

3.1. *The normal model*

Suppose that given θ_i, $(T_{1ij}, S_{1ij}, T_{2ij}, S_{2ij})^{\mathrm{T}}$ is normally distributed with mean θ_i and covariance matrix $\mathrm{diag}(\Sigma_{11i}, \Sigma_{22i})$. Then $\widehat{\theta_i} = (T_{1i}, S_{1i}, T_{2i}, S_{2i})^{\mathrm{T}}$. The quantity T_{1i} denotes the mean $n_i^{-1} \sum_{j=1}^{n_i} T_{1ij}$, with S_{1i}, T_{2i} and S_{2i} denoting similar averages over j. For $z = 1, 2$, we can estimate Σ_{zzi} as the sample covariance matrix of the terms (T_{zij}, S_{zij}), as indicated in Section 2. If the treatment effect of interest is $\delta_N = \theta_{1TN} - \theta_{2TN}$, then a confidence interval for δ_N given $\widehat{\theta_{SN}} = (\widehat{\theta_{1SN}}, \widehat{\theta_{2SN}}) = (S_{1N}, S_{2N})$ is given by $M(\Psi) \pm 1.96 V^{1/2}(\Psi)$, where $M(\cdot)$ and $V(\cdot)$ are given in (3) and (4). Because Ψ must be estimated, the plug-in prediction interval $M(\widehat{\Psi}) \pm 1.96 V^{1/2}(\widehat{\Psi})$ will have sub-nominal coverage. See Section 4 for a bootstrap procedure that yields a wider confidence interval with nominal coverage.

The confidence interval based on (3) and (4) is like those proposed by BMBRG for linear mixed models, except that they assume $\Sigma_{11i} = \Sigma_{22i}$. Daniels and Hughes (1997) base a confidence interval on the conditional distribution of δ_N given $S_{1N} - S_{2N}$. Estimation of δ_N based on (S_{1N}, S_{2N}) is more efficient, at least if Ψ is known, but the gains in efficiency are small in the following numerical examples.

Let the elements of ϕ be denoted by $\phi_{k\ell}$. To illustrate these ideas for scalar (T, S), let $\phi_{11} = 1$, $\phi_{22} = 4$, $\phi_{33} = 0.1$, and $\phi_{44} = 0.4$, and let the upper triangle of the correlation matrix for ϕ have first row $(1.0, 0.9, 0.7, 0.6)$, second row $(1.0, 0.6, 0.7)$, third row $(1.0, 0.9)$ and fourth row 1.0. Thus θ_{1S} and θ_{2S} are strongly correlated with θ_{1T} and θ_{2T} respectively, and S should be an excellent surrogate. Assume that a large number of previous studies permits us to estimate μ and ϕ precisely. Without data on (S_{1N}, S_{2N}), we would assert that $\theta_{1TN} - \theta_{2TN}$ is normal with mean $\mu_{1T} - \mu_{2T}$ and variance $\phi_{11} - 2\phi_{13} + \phi_{33} = 0.6573$. Now assume in addition that the study of the new drug is large, so that $(\sigma_{22N}, \sigma_{44N})$ are negligible compared to the elements of ϕ. Then from (3), the conditional expectation of $\theta_{1TN} - \theta_{2TN}$ given $(S_{1N}, S_{2N}) = (\hat{\theta}_{1SN}, \hat{\theta}_{2SN})$ reduces to $\mu_{1T} - \mu_{2T} + 0.4799(S_{1N} - \mu_{1S}) + 0.5636(S_{2N} - \mu_{2S})$; the residual variance about this predictor is 0.0880, and the fraction of variance explained is $(0.6573 - 0.0880)/0.6573 = 0.866$. BMBRG define this proportion of variance explained as the coefficient of determination R^2_{trial}, and suggest it as a figure of merit for a surrogate at the trial level.

In this example, had we used the regression on $S_{1N} - S_{2N}$ instead, as in Daniels and Hughes (1997), we would obtain the estimate $\mu_{1T} - \mu_{2T} + 0.4644(S_{1N} - \mu_{1S} + S_{2N} - \mu_{2S})$, with residual variance 0.0902 and fraction of variance explained 0.863. As one would expect, the fraction of variance explained decreases rapidly as the correlations between $(\theta_{1T}, \theta_{2T})$ and $(\theta_{1S}, \theta_{2S})$ diminish. Using 0.7 for these correlations in the previous example, instead of 0.9, we find that the fraction of variance explained drops to 0.398 for (3) and to 0.350 for the regression on $S_{1N} - S_{2N}$.

Even a good surrogate in the meta-analytic framework is not nearly as efficient as direct data on (T_{1N}, T_{2N}). Arguments similar to those above show that given $(T_{1N}, T_{2N}, \mu, \phi, \sigma_{11N}, \sigma_{33N})$, $\theta_{1TN} - \theta_{2TN}$ is normal with mean and variance similar in form to (3) and (4), where σ_{11N} and σ_{33N} are submatrices of Σ_{11N} and Σ_{22N} corresponding to $\hat{\theta}_{1TN}$ and $\hat{\theta}_{2TN}$ respectively. As the sample size of experiment N increases, σ_{11N} and σ_{33N} converge to zero, and the estimate of $\theta_{1TN} - \theta_{2TN}$ reduces to $T_{1N} - T_{2N}$ with a variance that tends to zero. Data from a surrogate are qualitatively weaker because as the sample size of experiment N increases, equation (4) tends to a value greater than zero. Indeed, no matter how large K is and no matter how large the new study is, one may be left with an irreducible residual variance that is unacceptably large.

3.2. Dichotomous outcomes

Let T_{zij} be the clinically relevant dichotomous response on subject j, study i, treatment arm z, and let S_{zij} be the corresponding dichotomous surrogate. The pairs (T_{zij}, S_{zij}) follow a multinomial distribution with index 1.0 and parameters $\pi_{zi} = (\pi_{zi11}, \pi_{zi10}, \pi_{zi01}, \pi_{zi00})$ corresponding to the outcomes (1, 1), (1, 0), (0, 1), and (0, 0).

We transform to parameters θ that might plausibly have a multivariate normal distribution by setting $\theta_{zTi} = \log\{(\pi_{zi11}+\pi_{zi10})/(\pi_{zi01}+\pi_{zi00})\}$ and $\theta_{zSi} = \log\{(\pi_{zi11}+\pi_{zi01})/(\pi_{zi10}+\pi_{zi00})\}$. Note that these parameters are the logarithms of marginal odds for T and S under treatment z. The two multinomials depend additionally on nuisance parameters $\gamma_i = [\log\{\pi_{1i11}\pi_{1i00}/\pi_{1i10}\pi_{1i01}\}, \log\{\pi_{2i11}\pi_{2i00}/\pi_{2i10}\pi_{2i01}\}]$ that we do not need to estimate.

Assuming θ_i is normally distributed, the estimate $\widehat{\theta_i}$ with components such as $\widehat{\theta}_{1Ti} = \log\{T_{1i}/(1-T_{1i}\}$ is normal with mean μ and covariance matrix $\phi + \Sigma_i$, as in Section 2. The covariance matrix Σ_i can be estimated as in Section 2 by setting $U_{1Tij} = T_{1ij} - \exp(\theta_{1Ti})\{1+\exp(\theta_{1Ti})\}^{-1}$ and defining U_{1Sij}, U_{2Tij} and U_{2Sij} similarly. Note that although the nuisance parameters γ_i determine Σ_i in part, we can estimate Σ_i without estimating γ_i.

If the estimated treatment effect is the log odds ratio $\delta_N = \theta_{1TN} - \theta_{2TN}$, (3) and (4) can be used to construct a confidence interval of the usual form, with modification in Section 4 to accommodate estimation of Ψ. Suppose instead we are interested in the risk difference $\delta_N = \exp(\theta_{1TN})/\{1 + \exp(\theta_{1TN})\} - \exp(\theta_{2TN})/\{1 + \exp(\theta_{2TN})\}$. If Ψ were known, we could simulate the conditional distribution of δ_N given $\hat{\theta}_{SN}$ by taking samples from the conditional distribution of θ_{TN} with mean (1) and variance (2). A confidence interval for δ_N could be based on this simulated distribution. Adaptations for estimation of Ψ are described in Section 4.

3.3. Marginal models for survival data and other complex data structures

Suppose T represents the time to death following cancer treatment and S the time to an event such as cancer recurrence. Note that S does not censor T, but if T precedes S, we assume that S is randomly censored at that time. Let $F(t|\theta_{zTi})$ and $F(s|\theta_{zSi})$ be the marginal distributions in experiment i on treatment z of T and S, respectively. For example, T_{zij} might have a Weibull distribution $\mathrm{pr}(T_{zij} \leq y) = 1 - \exp(-\lambda_{zTi} y^{\alpha_{zTi}})$. The parameterization $\theta_{zTi} = (\ln(\lambda_{zTi}), \alpha_{zTi})^{\mathrm{T}}$ and $\theta_{zSi} = (\ln(\lambda_{zSi}), \alpha_{zSi})^{\mathrm{T}}$ might plausibly conform to the multivariate normal distribution. These parameters define only the marginal distributions of T and S, not their joint distributions. We do this to avoid the complexity and difficulty of specifying and validating such joint distributions.

Estimates of these parameters are obtained by solving marginal score equations. The methods of Section 2 then apply immediately. In this example, the treatment effect might be expressed by the difference in median survivals $\delta_N = \{\ln(2)/\lambda_{1TN}\}^{\alpha_{1TN}} - \{\ln(2)/\lambda_{2TN}\}^{\alpha_{2TN}}$. This is not a linear function of θ_{1TN} and θ_{2TN}. Nonetheless, if Ψ were known, the distribution of δ_N given (S_{1N}, S_{2N}) could be obtained by simulating values of θ_{TN} given $\widehat{\theta}_{SN}$ from (1) and (2). In Section 4 we describe adaptations for estimated Ψ. This formulation requires that any censoring of T_{1ij} be independent of T_{2ij}, and similarly for S_{1ij} and S_{2ij}, but the four censoring distributions need not be the same.

Extending this example, one can use piecewise exponential models in which θ_{zTi} and θ_{zSi} are vectors of log hazard rates. In this context, δ_N might be the difference in median survival, the difference in estimated 5-year survival, the ratio of 5-year cumulative hazards, a weighted average of interval-specific

Table 1. *Estimated coverage of $\theta_{1TN} - \theta_{2TN}$ for the Carroll–Ruppert bootstrap procedure in Section 4.1*

Number of studies, K	Estimated coverage[a]		Average bootstrap	$\sqrt{\text{var}(v)}$[b]
	Bootstrap procedure[c]	Assuming $v = 1.96$	estimate of $\bar{v}$	
5	0.960	0.643	7.72	3.88
10	0.930	0.656	4.04	0.65
25	0.950	0.825	2.44	0.31
50	0.948	0.904	2.20	0.10
100	0.943	0.923	2.04	0.003

[a] Based on 1000 independent simulations. [b] This is the square root of the average estimate of var(v), as described in Section 4.1. [c] $B = 100$.

hazard ratios, which estimates a relative hazard, or some other function of θ_{1TN} and θ_{2TN}.

Another interesting example with survival data would be to let S represent survival information up to 2 years of follow-up and let T represent uncensored survival data. Then θ_{zTi} could contain piecewise exponential parameters as before, and θ_{zSi} contains a subset of the parameters in θ_{zTi} that define survival over the first 2 years. In a new trial, data over 2 years on S can be used to estimate δ_N, the difference in 5-year survival rates.

Repeated measures can be handled in a similar way by letting θ_{zTi} be parameters that define the evolution of the mean true response over time, while θ_{zSi} are parameters that define the evolution of the mean surrogate response. Here δ_N could be the difference in mean true response at a fixed time point, or some other quantity such as the difference in slopes of the two mean functions.

These methods, which are analogous to the GEE approach of Liang and Zeger (1995), are valid for estimating equations other than those implied by marginal models. For example, a normal marginal model yields an estimating equation for the mean. This mean may be a useful summary statistic even if the data do not follow a normal distribution. Since the sample standard error, or a standard error obtained by sandwich methods, is 'robust', i.e. consistent, our procedures are robust to misspecification of the model, provided the estimated parameters make sense. Correct modeling of the distribution of θ_i is far more critical.

4. BOOTSTRAP CONFIDENCE INTERVALS

4.1. Linear treatment effect

In Sections 2 and 3, we ignored variability in estimates of μ, ϕ, σ_{22N}, and σ_{44N}, which collectively we called Ψ . The previous variance estimates or conditional distributions are therefore only justified when a large number of previous studies have been conducted and when the current study, N, is large, so that these quantities are known with high precision. We now present data showing that ignoring the variability in $\widehat{\Psi}$ can lead to seriously misleading inference (see Table 1). To solve this problem, we propose a bootstrap procedure to estimate the variance of the conditional distribution of $\theta_{1TN} - \theta_{2TN}$ in Section 2 that results from plugging in estimates of unknown components of Ψ, and we use the method of Carroll and Ruppert (1991) to construct a confidence interval for $\theta_{1TN} - \theta_{2TN}$. Later we present alternative bootstrap methods for a more general treatment effect $\delta_N(\theta_{1TN}, \theta_{2TN})$.

Each cycle of the bootstrap has two phases. First, if the original data had K previous studies with complete data, then in each bootstrap replication, K such studies are re-sampled with replacement from the previous studies. For each re-sampled study, a new set of data is obtained. In particular, if the ith study is re-sampled, a new set of n_i observations is obtained by re-sampling treated subjects with replacement, and a new set of m_i observations is obtained by re-sampling control subjects with replacement. We also resample the observations in the current study, N, in the same way to obtain new estimates of σ_{22N} and σ_{44N}. Suppose we obtain B new bootstrap data sets in this way. For the bth set, we obtain the estimates of these parameters, which collectively we call $\widehat{\Psi}_b$. Holding $\widehat{\theta}_{SN}$ fixed, we obtain the plug-in estimate M_b from equation (3) and plug-in variance estimate V_b from equation (4). The variance of the conditional distribution of $\theta_{1TN} - \theta_{2TN}$ minus its plug-in estimate is the expectation over $(\widehat{\mu}, \widehat{\phi}, \widehat{\sigma}_{22N}, \widehat{\sigma}_{44N})$ of the conditional variance (4) plus the variance over $(\widehat{\mu}, \widehat{\phi}, \widehat{\sigma}_{22N}, \widehat{\sigma}_{44N})$ of the conditional expectation (3) with estimated parameters plugged in. These two components of variance can be estimated, respectively, by $B^{-1} \sum_b V_b$ and by $(B-1)^{-1} \sum_b (M_b - \overline{M})^2$.

In order to obtain a confidence interval on $\theta_{1TN} - \theta_{2TN}$, we use the method of Carroll and Ruppert (1991). Because $\theta_{1TN} - \theta_{2TN}$ is independent of $\widehat{\Psi}$, we can compute the expectation:

$$\begin{aligned} E\{H(\widehat{\Psi}, \Psi, v)\} \equiv\ & E(\Phi[\{M(\widehat{\Psi}) - M(\Psi) + vV^{1/2}(\widehat{\Psi})\}V^{-1/2}(\Psi)]) \\ & -E(\Phi[\{M(\widehat{\Psi}) - M(\Psi) - vV^{1/2}(\widehat{\Psi})\}V^{-1/2}(\Psi)]) = 1 - \alpha \end{aligned} \tag{5}$$

for some non-negative v, where Φ is the standard normal distribution function, $M(\widehat{\Psi})$ is the mean given by equation (3) with $\widehat{\theta}_{SN}$ held fixed, $V(\widehat{\Psi})$ is the variance given by equation (4), and $1 - \alpha$ is the desired coverage of the confidence interval. The expectation is over the distribution of $\widehat{\Psi}$. The aim is to solve (5) for v to create the confidence interval $M(\widehat{\Psi}) \pm vV^{1/2}(\widehat{\Psi})$. The previous bootstrap can be used to estimate the desired v as the solution to

$$B^{-1} \sum_b H(\widehat{\Psi}_b, \widehat{\Psi}, v) = 1 - \alpha. \tag{6}$$

If K, n_N, and m_N are large, so that $\widehat{\Psi}$ has little variability, v is nearly 1.96 for $1 - \alpha = 0.95$. With small K, it can happen that $V(\widehat{\Psi}_b) \leq 0$, in which case we discard that bootstrap sample and replace it with another sample for which $V(\widehat{\Psi}_b) > 0$.

To determine how well this procedure works and to estimate the effect of uncertainty in $\widehat{\Psi}$ on the width of the prediction interval, compared to the case where Ψ is known, we conducted a simulation study based on the normal model (Section 3.1). To simplify calculations, we assumed that Σ_{11i} and Σ_{22i} were fixed and known for $i = 1, 2, \ldots K$ and for $i = N$. In fact, $\Sigma_i = \text{diag}(\Sigma_{11i}, \Sigma_{22i})$ equaled $[0.1, .1, 0, 0; 0.1, 0.2, 0, 0; 0, 0, 0.1, 0.09; 0, 0, 0.09, 0.1]$, where successive rows are separated by semicolons. Thus $\widehat{\mu}$ was the mean of the K vectors $\hat{\theta}_i = (T_{1i}, S_{1i}, T_{2i}, S_{2i})^{\mathrm{T}}$, and ϕ was estimated by subtracting known elements of Σ_{11i} and Σ_{22i} from the sample covariance of $\hat{\theta}_i$ based on the K previous studies. We chose $\phi = [1, 1.8, 0.7, 1.2;\ 1.8, 4, 1.2, 2.8;\ 0.7, 1.2, 1, 1.8;\ 1.2, 2.8, 1.8, 4]$, and $B = 100$.

To assess the performance of the bootstrap procedure in this case, we generated 1000 independent data sets for each $K = 5$, 10, 25, 50, and 100. The data on the K earlier studies were generated directly from the distribution for $\hat{\theta}_i$, and the data for $i = N$ were generated from the distribution of $(\theta_{1TN}, S_{1N}, \theta_{2TN}, S_{2N})^{\mathrm{T}}$. The MATLAB (1997) random number generator 'mvnrnd' was used to produce pseudo-normal variates.

The Carroll and Ruppert (1991) bootstrap procedure yielded coverage near the nominal 0.95 level even with $K = 5$ (Table 1). In contrast, an analysis that assumes Ψ is known and therefore uses $v = 1.96$ has empirical coverage substantially below 0.95 for $K = 5$, 10, 25, and 50. For $K = 5$, the coverage probability even with $v = 7.0$ was only 0.870 (data not shown), compared to a coverage probability of 0.960 for the Carroll-Ruppert procedure (Table 1). For $K = 100$, the estimated coverage with $v = 1.96$ was 0.923 with 95% confidence interval (0.846, 0.940). These data indicate that in most applications, where K is limited, variability of $\widehat{\Psi}$ must be taken into account to obtain valid confidence intervals.

To determine the average value of estimates of v with $B = 100$, and to see how variable the estimates of v would be for $B = 100$ for a given data set, we simulated 100 additional data sets for each $K = 5$, 10, 25, 50, and 100. For each data set, s, we repeated the bootstrap procedure (with $B = 100$) 100 times and computed the mean and sample variance of these 100 estimates of v for data set s. We then computed the average mean (called $\bar{v}$) and average sample variance ($\overline{\text{var}v}$) over $s = 1, 2, \ldots, 100$ (see Table 1). For $K = 100$, the average $\overline{v} = 2.04$ is not much greater than the nominal 1.96, whereas for $K = 5, \overline{v} = 7.72$ represents a huge loss of precision (Table 1). Thus, although the bootstrap procedure covers $\theta_{1tN} - \theta_{2tN}$ at nominal levels for $K = 5$, it does so by expanding the width of the prediction interval enormously. For $K = 5$, the ratio $7.72/1.96 = 3.9$ indicates the loss in precision from having to estimate Ψ.

With $B = 100$, there is considerable variability in the bootstrap estimate of v for small K, as indicated by $\sqrt{\overline{\text{var}v}}$ (Table 1). For $K = 100$, the ratio of $\sqrt{\overline{\text{var}v}}$ to $\overline{v}$ is $0.003/2.037 = 0.001$, whereas, for $K = 5$, this ratio is 0.502. This suggests that larger bootstrap samples should be used for small K, even though the coverage is adequate with $B = 100$. With $B = 1000$ and $K = 5$, we found that the ratio was only 0.104.

4.2. *General treatment effects*

Alternative methods are needed when δ_N is not a linear function of $(\theta_{1TN}, \theta_{2TN})$. We therefore propose the following parametric bootstrap procedure. Obtain bootstrap samples $\widehat{\Psi}_b$ for $b = 1, \ldots, B$ as described in Section 4.1. For each $\widehat{\Psi}_b$ draw an observation from the conditional distribution, (1) and (2), of θ_{TN} given $\widehat{\theta}_{SN}$ and $\widehat{\Psi}_b$, and then compute $\delta_{N,b}$. Combining these B values gives an estimate $\hat{G}_B$ of $G(\delta_N | \widehat{\theta}_{1SN}, \widehat{\theta}_{2SN}, \Psi)$, the conditional distribution function of δ_N given $\widehat{\theta}_{1SN}, \widehat{\theta}_{2SN}$ and Ψ. This approach takes into account the variability that is introduced by using $\widehat{\Psi}$ in place of Ψ when calculating $\hat{\delta}_{N,b}$. The confidence interval for δ_N is $\{\hat{G}_B^{-1}(\alpha/2), \hat{G}_B^{-1}(1 - \alpha/2)\}$.

To show that the method yields nominal coverage, we first tested it on $\delta_N = \theta_{1TN} - \theta_{2TN}$, using the parameters presented in Section 4.1. The coverage was within the sampling error of nominal levels, as for the Carroll–Ruppert procedure, but the confidence intervals were somewhat wider. For $K = 25$ previous studies, for example, the average length of the confidence intervals for the treatment differences was 1.51 for the parametric bootstrap and 1.40 for the Carroll–Ruppert bootstrap procedure. Both these procedures for $K = 25$ yielded confidence intervals about 25% longer than would be the case if Ψ were known.

We applied this procedure to obtain confidence intervals for the treatment effect $\delta_N = \exp(\theta_{1TN})/\{1 + \exp(\theta_{1TN})\} - \exp(\theta_{2TN})/\{1 + \exp(\theta_{2TN})\}$, defined for the binary outcome example (Section 3.2). To be specific, we choose $\pi_{1i11} = 0.6, \pi_{1i10} = 0.2, \pi_{1i01} = 0.05, \pi_{1i00} = 0.15$ and $\pi_{2i11} = 0.5, \pi_{2i10} = 0.1, \pi_{2i01} = 0.1, \pi_{2i00} = 0.3$. Application of the delta method and using $n_i = m_i = 1000$, gives the covariance matrices $\Sigma_{11} = [0.0625, 0.0219; 0.02198, 0.04396]$ and $\Sigma_{22} = [0.0416, 0.02431; 0.02431, 0.0416]$. The matrix ϕ was chosen to be the same as in the linear example.

Table 2 presents the average length of the confidence intervals for different numbers of previous studies, and shows that the parametric bootstrap yielded near nominal coverage for this nonlinear treatment effect, except for the case $K = 5$. In that case, the observed coverage, 0.908, fell below the range expected to cover 95% of the simulated results, (0.931, 0.969). This is not surprising, as the empirical distribution function based on so few observations is a poor estimate of the true G.

5. EXAMPLE

Our methods are illustrated using data from the Regression Growth Evaluation Statin Study (REGRESS) trial (see Jukema *et al.*, 1995), a placebo-controlled multicenter study to assess the effects of two years of treatment with pravastatin on progression and regression of coronary atherosclerosis in men scheduled for

Table 2. *Estimated coverage of* $\delta_N = exp(\theta_{1tN})/\{1 + exp(\theta_{1tN})\} - exp(\theta_{2tN})/\{1 + exp(\theta_{2tN})\}$*, for the parametric bootstrap procedure described in Section 4.2*

Number of studies, K	Estimated coverage[a]	Average length of the confidence intervals	Ratio of CI widths[b]
5	0.908	0.8201	5.26
10	0.934	0.5071	3.25
25	0.921	0.4176	2.68
50	0.948	0.3264	2.09
100	0.962	0.2545	1.63

[a] Based on 500 independent simulations, each with $B = 5000$ bootstrap samples. [b] This is the ratio of the width of the 95% confidence interval from the bootstrap procedure to that assuming that Ψ is known, namely 0.156.

arteriography who had a total cholesterol in the range 4–8 mmole/L (155–310 mg/dL). At baseline and after two years of treatment a coronary angiogram was made. Serum cholesterol was measured at baseline and during follow-up, and events were recorded. The primary endpoint, T_{ij}, is the change in average mean coronary artery segment diameter over the two-year trial period. The surrogate outcome, S_{ij}, is the change in serum cholesterol during follow-up. Histograms of T and S indicate that the differences can be assumed to arise from a bivariate normal distribution.

Although these data did not come from a series of independent clinical trials, as in a true meta-analysis, we treated the centers as if they were independent studies. The first 10 centers correspond to 'previous' studies with full information on the surrogate and primary endpoints. The eleventh center was chosen to represent the 'new' study, where only surrogate information is available.

After eliminating patients with missing information, there were 61, 32, 70, 48, 98, 67, 52, 21, 96, 58, and 32 subjects in each 'study'. Because the sample sizes in the 10 'previous studies' are small, and because the between-center variation in these data is not large, the estimate of ϕ, $\hat{\phi} = \widehat{\text{cov}}(T_{1i}, S_{1i}, T_{2i}, S_{2i}) - \frac{1}{K}\sum_{i=1}^{K} \hat{\Sigma}_i/n_i$ was not positive definite. To obtain more realistic sample sizes for the individual studies, we re-scaled the individual sample size so that on average they had the same size as the original trial. The resulting desired sample sizes were 670, 352, 770, 528, 1078, 736, 572, 230, 1056, 638, and 352 (Table 3). In order to obtain these larger 'samples', we estimated the individual-level center-specific and treatment-specific means and covariances. Assuming bivariate normality, we generated 22 different samples with equal numbers in the pravastatin and placebo groups in each 'study'. Table 3 presents the total sample sizes, the T_{1i}, S_{1i}, T_{2i} and S_{2i}, and the estimated correlation coefficients, based on the individual level data in each 'study' and treatment group.

The estimated treatment effects for the 'new' study (center 11) are $T_{1N} - T_{2N} = -0.0725 - (-0.1106) = 0.0381$ mm with 95% confidence interval $[-0.0138, 0.0900]$ based on the true responses in that study. The estimated treatment effect based on the surrogate data is 0.0402, with the naive confidence interval $[-0.0552, 0.1355]$ from equation (4) . The bootstrap procedure from equations (5) and (6) with $B = 500$ yields $\nu = 3.5910$ and the more realistic 95% confidence interval $[-0.1346, 0.2149]$. The width of this interval, 0.3495, is 3.3671 times greater than the width of the interval based on $T_{1N} - T_{2N}$, illustrating a very serious loss in efficiency from relying on the surrogate.

In the previous calculations, we let center 11 correspond to the 'new' study. When, instead, we chose center 5, which has the largest sample, to represent the 'new' study, the loss in efficiency comparing the surrogate approach to the true measurement $T_{1N} - T_{2N}$ was even more dramatic. The estimate based on the true response was 0.0074 with 95% confidence interval $[-0.0156, 0.0303]$. The estimate based on equation (3) was 0.0465 with bootstrap 95% confidence interval $[-0.2144, 0.3074]$, and the ratio of the widths of those two confidence intervals was 11.37.

Table 3. *Data based on treatment- and center-specific parameters estimated from the REGRESS trial*

Center	Total sample size	$(T_{1i}, S_{1i})^a$	(T_{2i}, S_{2i})	Individual-level correlation	
				Pravastatin	Placebo
1	670	(−0.0133, −1.3800)	(−0.1597, 0.0712)	0.0378	0.0863
2	352	(−0.0860, −1.3875)	(−0.0842, −0.0348)	−0.2817	−0.3352
3	770	(−0.0772, −1.2829)	(−0.1128, 0.1957)	−0.0102	−0.3044
4	528	(−0.0274, −1.2622)	(−0.1712, 0.2126)	−0.1807	0.6712
5	1078	(−0.0515, −1.2319)	(−0.0588, 0.2517)	0.0507	0.2116
6	736	(−0.1235, −1.0390)	(−0.0987, 0.3215)	−0.0885	0.2863
7	572	(−0.0694, −1.3673)	(−0.0960, 0.1006)	0.2023	0.1790
8	230	(0.0029, −0.9688)	(−0.0374, 0.0827)	−0.5844	0.3586
9	1056	(−0.0361, −1.3233)	(−0.0944, 0.0644)	−0.0780	−0.0650
10	638	(−0.1250, −0.9502)	(−0.1302, 0.0010)	−0.0181	−0.0224
11	352	(−0.0725, −1.2008)	(−0.1106, −0.0538)	0.0420	0.0069
Means	634.7	(−0.0617, −1.2176)	(−0.1049, 0.1103)	−0.0826	0.0975

[a] Units of T are mm and units of S are mmole/L.

The original study by Jukema *et al.* (1995) estimated the treatment effect as 0.04 mm with confidence interval $[0.01, 0.07]$ and significance $p = 0.019$. Note that if the surrogate marker had been used, realistic confidence intervals such as those above would have been much too wide to demonstrate a treatment effect on the primary outcome. Unfortunately one can not compensate effectively in the surrogate approach by increasing the size of the studies, because much of the variation is due to ϕ. Increasing the number of 'previous' studies can reduce the variability of the estimate of ϕ and improve precision somewhat. To take an extreme case, suppose there had been hundreds of previous studies, so that we can assume ϕ is known without error. Suppose also that the new study was very large so that σ_{22N} and σ_{44N} were negligible. Using the estimate of ϕ obtained from the first 10 studies in Table 3, we estimate the variance of the surrogate-based estimate of treatment effect from equation (3) as 0.0023, which corresponds to a width of the 95% confidence interval of 0.188 mm instead of 0.06 mm found in the original study. Thus, no matter how many previous studies were available and no matter how large the new study was, this surrogate would not yield a sufficiently precise estimate of treatment effect.

If we assume that ϕ is known without error, we can calculate the proportion of variance explained by the surrogate, illustrated in Section 3.1, as $(0.0028 - 0.0024)/0.0028 = 0.1476$. In this calculation, 0.0028 is the variance of the estimated treatment differences $\hat{\theta}_{1TN} - \hat{\theta}_{2TN} = \hat{\mu}_{1T} - \hat{\mu}_{2T}$ based on the 'previous' studies, but without any use of surrogate data, and 0.0024 is the smaller variance that results from using surrogate data in equation (3). If we assume in addition that the new trial is so large that $\sigma_{22N} = \sigma_{44N} = 0$, then the proportion of variance explained becomes $(0.0028 - 0.0023)/0.0028 = 0.1786$. This quantity is the coefficient of determination, R^2_{trial}, discussed by BMBRG.

One would hope that the changes in cholesterol would be strongly negatively correlated with the changes in arterial diameter. The correlation between the 11 values of $T_{1i} - T_{2i}$ and $S_{1i} - S_{2i}$ in Table 3 is −0.3049. Note that the center- and treatment-specific individual-level correlations are variable and are, on average, much smaller than 0.3049 in absolute value. This variation indicates that cholesterol change is not a reliable indicator of arterial diameter change at the individual level, even though it has some predictive value at the trial level.

The ϕ in the present example is probably not typical of the between-study variation that one would find in a true meta-analysis covering a variety of drugs and studies. The dispersion of θ_i in the present study is reduced because only a single agent is being evaluated and all the study centers are adhering to the same protocol.

6. Discussion

This work was motivated by the studies of Daniels and Hughes (1997) and Buyse *et al.* (2000), who used the meta-analytic approach. This paper extends the work of BMBRG by introducing methods for a very general class of models that only require that summary parameters of interest be specified (Section 2). These methods could be applied, for example, to analyze separate piecewise exponential survival curves for T and S in treated and untreated groups. Moreover, these method are easily extended to allow the surrogates to be vector valued, even if T is a scalar. In the special case of the normal model (Section 3.1), we allow a more general covariance structure for Σ_{ii} than that used by BMBRG. Unlike earlier work, our study also indicates the critical importance of taking into account the uncertainty in estimates of the parameters that govern the distribution of θ_i (Section 4), and we introduce bootstrap methods to take such variability into account.

The meta-analytic approach relies very little on modeling the relationship between T and S at the individual level. Instead, inference is based on the empirical distribution, in a series of previous studies, of summary parameters, such as group means, that characterize the responses T and S in treated and untreated groups. In this paradigm, the ability to predict the treatment effect of a new drug depends principally on how tightly summary parameters for T are related to summary parameters for S in other studies of drugs from the same class.

BMBRG proposed a coefficient of determination based only on elements of ϕ to measure the adequacy of a surrogate. This trial-level criterion can be extended to cover other measures of treatment effect, δ_N, such as effects on a logit scale, by calculating the fractional reduction in trial-level variance of the estimate of δ_N that results from using $\hat{\theta}_{1SN}$ and $\hat{\theta}_{2SN}$. In applications, however, the surrogate will not perform as well as indicated by the coefficient of determination or its extension because ϕ must be estimated, which can seriously degrade the precision of estimates based on the surrogate (Section 4). A more realistic estimate of the reduction in variance from using $\hat{\theta}_{1SN}$ and $\hat{\theta}_{2SN}$ could be based on the bootstrap (Section 4). The bootstrap estimate of the variance of $\hat{\delta}_N$ based only on data from previous studies can be compared to the bootstrap estimate of the variance of $\hat{\delta}_N$ based on both the previous studies and on $\hat{\theta}_{1SN}$ and $\hat{\theta}_{2SN}$. In these calculations, one would assume $\Sigma_N = 0$.

An alternative approach to estimating treatment effects could be based on strong assumptions about the ability of S to predict T at the individual level. In an ideal case, T would be conditionally independent of any treatment given S (Prentice, 1989). Such an assumption is essential for proving that hypothesis tests of no treatment effect based on S are valid for testing for treatment effects on T (Prentice, 1989). Moreover, if we assume that T is conditionally independent of any treatment and of any other factors, given S, and if the conditional distribution $F(t|s)$ has been estimated, for example from untreated subjects, then for any other treatment, $Z = z$, the relevant marginal distribution can be calculated from surrogate data on $F(s|z)$ via $F(t|z) = \int F(t|s)dF(s|z)$. Such a strong assumption would need to be justified by a thorough understanding of the biological relationship between S and T, and by assurance that no factor influences T except through its influence on S. Even if several previous studies indicated little dependence of $F(t|s, z)$ on z, there remains a possibility that a new drug might influence this relationship (Fleming and DeMets, 1996). To relax this strong assumption, some researchers have turned to the concept of percentage of treatment effect explained by the surrogate (Freedman *et al.*, 1992; Lin *et al.*, 1997; Buyse and Molenberghs, 1998), but this quantity does not allow one to predict the effect of a new treatment on T from data on its effect on S. The meta-analytic approach avoids dependence on the assumption of conditional independence at the individual level by relying, instead, on an empirical evaluation of group-level summary statistics.

Meta-analytic calculations in this paper can be used to determine whether a new study based on S can yield sufficiently precise estimates of the treatment effect on T to be useful. In particular, a prediction interval based on (2) or (4) may be too wide, and the corrected prediction interval, which takes the uncertainty of plug-in estimates into account, will be even wider (Section 4).

We have glossed over a complication in Sections 2 and 3. The variables (T_{zij}, S_{zij}) depend not only on θ_i but also on other nuisance parameters, γ_i (Figure 1). For example, in Section 3.1, γ_i, n_i and m_i define the covariance Σ_i. This covariance can be estimated without estimating γ_i, however. Likewise, Σ_i is estimated in Section 3.2 without estimating the nuisance parameters. In the class of problems discussed in Sections 2 and 3, the parameters of the asymptotic normal distribution for $\hat{\theta}_i$ can be estimated without knowledge of γ_i, though, in fact, the distribution of $\hat{\theta}_i$ is conditional on the unknown γ_i. Thus, inference on δ_N can be carried out as in Sections 2 and 4 without estimating γ_i.

In Sections 2 and 3 we have assumed that the estimating equation for θ_{1Ti} is functionally independent of $\theta_{1Si}, \theta_{2Ti}, \theta_{2Si}$ and γ_i, and other estimating equations likewise depend only on a single component of θ_i. If one wants to make additional assumptions on θ_i, such as a proportional hazards assumption between the parameters θ_{1Ti} and θ_{2Ti}, which might correspond to piecewise exponential log-hazards, for example, then the estimating equations might depend on more than one set of parameters, such as θ_{1Ti} and θ_{2Ti}. The procedures in Section 2 can be generalized to cover this case, although Σ_i would no longer necessarily be block- diagonal. Assuming Σ_i can be estimated by an extension of the methods in Section 2, however, the formulas in the paper would remain unchanged, except that $W\Sigma_N W^{\mathrm{T}}$ is no longer block-diagonal, because it must also include within experiment covariance matrices $\sigma_{24N} = \mathrm{cov}(\widehat{\theta_{1SN}}, \widehat{\theta_{2SN}})$, and $\Psi = (\mu, \phi, \sigma_{22N}, \sigma_{24N}, \sigma_{44N})$.

Although the meta-analytic approach offers an empirical basis for estimation, several challenges must be overcome in order for this approach to be useful. Indeed, some of these difficulties may seriously degrade the validity and utility of this approach.

First, it is difficult to define the class C of 'similar' drug studies to which the current study N belongs. Presumably, the new drug has the same supposed mechanism of action as other drugs in C, and the same types of control treatments are used in all studies. However, the underlying parameters θ_i will vary from study to study in class C not only because the drugs differ, but also because the populations studied differ. Critics may dispute the definition of C and the previous studies included in the analysis. Such criticism is common in meta-analyses to combine data on main endpoints.

Second, there may have been too few previous studies in C with complete data on T and S to permit reliable estimation of the distribution governing parameters θ_i. This can have a very deleterious impact on the precision of the estimated treatment effects (Section 4). Indeed, the results in Tables 1 and 2 indicate that prediction intervals that properly take uncertainty of $\widehat{\Psi}$ into account are much wider than naive prediction intervals based on the assumption that $\widehat{\Psi}$ is known when K is 10 or less. More simulations are needed to examine the range of parameters over which such conclusions hold, but these findings resemble those for empirical Bayes confidence intervals (Laird and Louis, 1987).

Third, it may be difficult to obtain individual-level data from previous studies that are needed to estimate Σ_{11i} and Σ_{22i}. This may be a matter of secondary importance if between-study variation is much greater than within-study variation (see Daniels and Hughes, 1997).

Fourth, the precision of estimated treatment effects from the meta-analytic approach is limited by between-study variation in the parameters θ_i in C, or, in other words, by the nature of ϕ, even when Ψ is known. In situations with substantial between-study variation in $(\theta_{1Ti}, \theta_{2Ti})$, even a surrogate that is strongly correlated with T can yield much less precise estimates of treatment effects on T than estimates based on the main endpoint T itself (Sections 3.1 and 4). No matter how many studies one includes in the meta-analysis, and no matter how large the new study is, the use of a surrogate may result in an irreducible variance of the estimated treatment effect, governed by ϕ, that is unacceptably large.

Fifth, meta-analytic approaches require realistic models for the distribution of θ_i. The multivariate normal model in Section 3.1 was chosen for mathematical convenience, but it may offer a reasonable approximation after transformation to a suitable set of parameters θ; see, e.g., Section 3.2. Other models could be used but the analysis of such models could be quite complicated and would often require computer-intensive techniques, such as Markov chain Monte Carlo (Daniels and Hughes, 1997). Nonetheless, an important issue for the meta-analytic approach is the sensitivity of the inference to modeling the distribution of θ_i.

Sixth, there is considerable need for research on methods to apply the meta-analytic approach to other types of data, such as data on repeated measures and survival data. Although the ideas in Section 2 and 3 may be useful in these more complex settings, there are unresolved issues in defining the surrogate and in modeling joint or marginal distributions of T and S. There is also a need to define the strengths and weaknesses of frequentist, empirical Bayes, and Bayesian approaches for analysis of the hierarchical system. Other improvements, such as the use of covariates to adjust for between-study variation in θ_i also warrant study.

Finally, there is the problem of unanticipated delayed toxicity that is not encompassed by the main endpoint T. For example, suppose T is the time to cancer recurrence and S is the initial degree of tumor shrinkage following cancer treatment. A study that is long enough to measure time to recurrence affords a greater opportunity to detect an unanticipated delayed toxicity than a shortened study based on the surrogate. Moreover, assessment of a new treatment may require evaluating a variety of true endpoints so that T becomes a vector. In these cases, the use of surrogates becomes more complex and less attractive.

Despite these many potential obstacles, the meta-analytic approach warrants further methodological study and efforts at practical implementation. Such studies will determine whether the approach has clinical utility. Even under propitious circumstances, however, our calculations (Sections 3.1 and 4) indicate that meta-analyses of surrogates will lead to much less precise estimates of treatment effect on T than relying on true endpoints, and the difficulty of defining an appropriate class of 'similar' studies will also make reliance on surrogates less convincing.

Acknowledgements

We are grateful to the editor and reviewers for helpful comments and to Dr Koos Zwinderman and other investigators and participants in the REGRESS trial for use of the data in Section 5. Carroll's research was supported by a grant from the National Cancer Institute (CA-57030) and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institutes of Environmental Health Sciences (P30-ES09106). van Houwelingen's research was conducted during a visit to the Biostatistics Branch of the Division of Cancer Epidemiology and Genetics, National Cancer Institute.

References

Buyse, M. and Molenberghs, G. (1998). Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics* **54**, 1014–1029.

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D. and Geys, H. (2000). The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*, **1**, 49–67.

Carroll, R. J. and Ruppert, D. (1991). Prediction and tolerance intervals with transformation and/or weighting. *Technometrics* **33**, 197–210.

CARROLL, R. J., RUPPERT, D. AND STEFANSKI, L. A. (1995). *Measurement Error in Nonlinear Models.* London: Chapman and Hall.

DANIELS, M. J. AND HUGHES, M. D. (1997). Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine* **16**, 1965–1982.

FLEMING, T. R. AND DEMETS, D. L. (1996). Surrogate endpoints in clinical trials: are we being misled? *Annals of Internal Medicine* **125**, 605–613.

FREEDMAN, L. S., GRAUBARD, B. I. AND SCHATZKIN, A. (1992). Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine* **11**, 167–178.

JUKEMA, J. W., BRUSCHKE, A. V. G., VAN BOVEN, A. J., REIBER, J. H. C., BAL, E. T., ZWINDERMAN, A. H., JANSEN, H., BOERMA, G. J. M., VAN RAPPARD, F. M. AND LIE, K. I. (1995). Effects of lipid lowering by pravastatin on progression and regression of coronary artery disease in symptomatic men with normal to moderately elevated serum cholesterol levels. The Regression Growth Evaluation Statin Study (REGRESS). *Circulation* **91**, 2528–2540.

LAIRD, N. M. AND LOUIS, T. A. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *Journal of the American Statistical Association* **82**, 739–750.

LIANG, K. Y. AND ZEGER, S. L. (1995). Inference based on estimating functions in the presence of nuisance parameters. *Statistical Science* **10**, 158–173.

LIN, D. Y., FLEMING, T. R. AND DEGRUTTOLA, V. (1997). Estimating the proportion of treatment effect explained by a surrogate marker. *Statistics in Medicine* **16**, 1515–1527.

MATLAB (1997). *Statistics Toolbox User's Guide*. Natick, MA: The Math Works.

PRENTICE, R. L. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine* **8**, 431–440.

[*Received October 15, 1999*; *revised March 9, 2000*; *accepted for publication March 22, 2000*]